Segmenting bacterial and viral DNA sequence alignments with a transdimensional phylogenetic factorial hidden Markov model

Wolfgang P. Lehrach†

Biomathematics and Statistics Scotland, Edinburgh, UK and Institute of Adaptive and Neural Computation, University of Edinburgh, UK.

Dirk Husmeier

Biomathematics and Statistics Scotland, Edinburgh, UK.

Abstract. The traditional approach to phylogenetic inference assumes that a single phylogenetic tree can represent the relationships and divergence amongst the taxa. However, taxa sequences exhibit varying levels of conservation, e.g. due to regulatory elements and active binding sites. Also, certain bacteria and viruses undergo interspecific recombination, where different strains exchange or transfer DNA subsequences, leading to a tree topology change. We propose a phylogenetic factorial hidden Markov model to simultaneously detect recombination and rate variation. This is applied to three DNA sequence alignments: one bacterial (Neisseria), the second of HIV-1, and the third from a family of plan actin genes. Inference is carried out in the Bayesian framework, using RJMCMC.

Keywords: Bayesian; HIV-1; Neisseria; Phylogenetic Factorial HMM; Reversible Jump MCMC

1. Introduction

The underlying assumption of most phylogenetic reconstruction methods is that a single phylogenetic tree captures the evolutionary history of a set of taxa. A phylogenetic tree is defined by its topology and its branch lengths. The topology (branching order) defines the way the taxa are related. The branch lengths indicate the average mutational divergence between the associated taxa. These trees are generally estimated from an alignment of DNA or protein sequences, where the corresponding DNA or protein sequence is taken from each taxa. However in functional regions of proteins, such as the binding site for oxygen in haemoglobin or catalytically active sites in enzymes, the average divergence between the sequences decreases (Nimrod et al., 2005). Mutations in these areas are likely to adversely affect the probability of the organism surviving until reproduction, and thus these mutations are less likely to become fixed in the population. Hence, these differences in the divergence indicate areas of interest along the alignment, which is exploited in fields such as comparative genomics to find conserved regulatory elements (Chen and Blanchette, 2007). These variations in the conservation rate along the sequence alignment cannot be captured when only a single phylogenetic tree with fixed branch lengths is used to represent the evolutionary relationships and divergences among the taxa. Furthermore, while the assumption of an unchanging hierarchy between the taxa is reasonable when applied to most DNA sequence alignments, it can be violated in certain bacteria and viruses due to

†Address for correspondence: 5 Forrest Hill, Edinburgh EH1 2QL, UK. Email: wlehrach@ed.ac.uk

interspecific recombination. The resulting transfer or exchange of DNA subsequences can lead to a change of the branching order (topology) in the affected region, which results in conflicting phylogenetic information from different regions of the alignment. If undetected, the presence of these so-called mosaic sequences can lead to systematic errors in the estimation of the phylogenetic tree and the rate of divergence along the sequence (Husmeier and Wright, 2001). Their detection, therefore, is a crucial prerequisite for consistently inferring the evolutionary history of a set of DNA sequences.

Various methods for detecting evidence of interspecific recombination in DNA sequence alignments have been developed; see, for instance, Husmeier et al. (2005) for a recent review. The objective of the present article is to discuss how the performance of simultaneously detecting rate variation and recombination using a recently proposed (Husmeier, 2005) combination of phylogenetic trees with Hidden Markov models (HMMs) can be substantially improved.

HMMs provide a powerful tool widely used in Bioinformatics (Baldi and Brunak, 1998), and they have been successfully applied to the segmentation of DNA sequences (Boys et al., 2000; Boys and Henderson, 2001, 2004). Here, the objective is to locate homogeneous segments within individual DNA sequences which are compositionally different from the rest of the sequence. The hidden states represent the homogeneous segments to be detected, which are characterised by their distribution of nucleotides, or by their first-order Markovian transition probabilities between nucleotides (Boys et al., 2000). A critical question is to infer how many different segment types a DNA sequence is composed of. To this end, Boys and Henderson (2001, 2004) adopted a Bayesian approach and sampled the number of hidden states from the respective posterior distribution with reversible jump (RJ) Markov chain Monte Carlo (MCMC).

The problem of detecting recombination and rate variation is related to the segmentation of a single sequence described above, but differs from it in two important aspects. First, the data to be segmented is not a single DNA sequence but a DNA sequence alignment. Second, homogeneity in a segment is not defined with respect to the nucleotide composition, but with respect to the underlying evolutionary history. This evolutionary history is captured by a phylogenetic tree, consisting of its topology H_S and associated vector of branch lengths $\mathbf{w}$ (depending on the nucleotide substitution model used, there might be some additional model dependent parameters $\boldsymbol{\theta}$). Hence, in generalisation of both the standard HMM applied to DNA sequence segmentations and the traditional approach to phylogenetics, one can marry the HMM to a phylogenetic tree – henceforth referred to as a phylogenetic HMM – where the latter defines the emission probabilities associated with the columns in the alignment.

Phylogenetic HMMs were originally introduced by Felsenstein and Churchill (1996) to allow for correlations between evolutionary rates at different sites. The rates associated with the hidden states were set to *a priori* fixed values that were not inferred from the data. Siepel and Haussler (2004) applied phylogenetic HMMs to model mosaic structures in DNA sequence alignments in the context of comparative genomics. The parameters were inferred by maximum likelihood in a supervised way, assuming that the hidden state sequences were known. The application of phylogenetic HMMs to the detection of recombination was first proposed by McGuire et al. (2000), with subsequent improvements of the inference methodology by Husmeier and Wright (2001) and Husmeier and McGuire (2003). However, these models can confuse regions subject to recombination and rate variation. Husmeier (2005) addressed this problem by introducing a phylogenetic factorial HMM (FHMM – see Ghahramani and Jordan, 1997), with two different types of hidden states. This disentangles topology changes – indicative of recombination – from changes of the nucleotide substitution

rate. For the latter, a set of fixed, *a priori* chosen values was used, akin to the approach of Felsenstein and Churchill (1996). This set of fixed rates limits the accuracy to which the rate variation along the sequence can be characterised.

The model proposed in this paper improves on the approach of Husmeier (2005) in three important respects. First, rather than setting the parameters associated with the hidden states to *a priori* selected fixed values, we place a prior distribution on them. Second, we infer the number of hidden states, which corresponds to the number of homogeneous segments in the DNA sequence alignment, with RJMCMC. Finally, we also apply this inference scheme to allow for changes in the transition-transversion ratio along the alignment.

An alternative approach to the HMM is the Multiple Changepoint Model of Suchard et al. (2003) and Minin et al. (2005). Here, the tree topology, the rate, and the transition-transversion ratio are allowed to vary between change points. The number and location of change points are inferred in a Bayesian framework and sampled from the posterior distribution with RJMCMC. We show that this model can be interpreted as a special case of a phylogenetic FHMM, and we compare the performances of both approaches empirically.

The article is organised as follows. We describe the model in Section 2, while Section 3 outlines our MCMC inference scheme. Section 4 compares our model with the breakpoint models of Suchard et al. (2003) and Minin et al. (2005), while Section 5 discusses an alternative model with coupled estimation of the rate and transition-transversion ratio. In Section 6 we discuss our predictions on a bacterial DNA sequence alignment, in Section 7 we examine some predictions on an alignment of maize actin genes, and Section 8 investigates the predictions for a DNA sequence alignment of Human Immunodeficiency Virus type 1 (HIV-1). Section 9 contains a discussion of our work, and Section 10 is a concluding summary.

2. The Model

2.1. The Bayesian phylogenetic factorial hidden Markov model (FHMM)

Consider an alignment $\mathcal{D}$ of m DNA sequences, N nucleotides long. Let a column in the alignment be represented by $\mathbf{y}_t$, where the subscript t represents the site, $1 \leq t \leq N$. Hence $\mathbf{y}_t$ is an m-dimensional column vector that contains the nucleotides at the t^{th} site of the alignment, and $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. The traditional approach to phylogenetics assumes that sites in the DNA sequence alignment are identically and independently distributed (iid); see, for instance, Durbin et al. (1998) or Husmeier et al. (2005) for a review. Given a tree topology H_S (the notation will become clear later), an associated vector of branch lengths $\mathbf{w}$, and a nucleotide substitution model (with extra parameters $\boldsymbol{\theta}$), the probability of the DNA sequence alignment is given by:

$$P(\mathcal{D}|H_S, \mathbf{w}, \boldsymbol{\theta}) \;=\; \prod_{t=1}^{N} P(\mathbf{y}_t|H_S, \mathbf{w}, \boldsymbol{\theta}). \tag{1}$$

$P(\mathbf{y}_t|H_S, \mathbf{w}, \boldsymbol{\theta})$ denotes the probability of the t^{th} column in the alignment, and is defined by the nucleotide substitution model used. In this paper, we use HKY85, the reversible Markov process model introduced by Hasegawa et al. (1985), which has the nucleotide substitution

rate matrix $\mathbf{N}$:

$$\mathbf{N} = \begin{pmatrix} - & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & - & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & - & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & - \end{pmatrix}, \tag{2}$$

where π_A, π_C, π_G and π_T are the equilibrium probabilities of the nucleotides, α and β are the transition and transversion rates, and the four rows and columns of the matrix refer to the four nucleotides in the order thymine (T), cytosine (C), adenine (A), and guanine (G). The diagonal elements are given by the constraint that each row of the matrix sums to zero. If $S(t)$ is the distribution over the nucleotides at time t and h is a sufficiently short time interval, then $P(S(t+h) = j | S(t) = i) = \mathbf{N}_{i,j} + o(h^2)$ where $j \neq i$ – the entries of $\mathbf{N}$ are the instantaneous transition probabilities.

Dividing $\mathbf{N}$ by β allows $\mathbf{N}$ to be expressed in terms of $\kappa = {}^{\alpha}/_{\beta}$. Instead of using κ like Husmeier (2005) and Minin et al. (2005), we follow the DNAML (Felsenstein, 1981) and PUZZLE (Schmidt et al., 2002) approach and use $E = \kappa \frac{(\pi_T\pi_C + \pi_A\pi_G)}{(\pi_T+\pi_C)(\pi_A+\pi_G)}$, the ratio of expected transition mutation events to transversion mutation events (Rosenberg et al., 2003), which we will refer to as the transition-transversion ratio. We define $\tau = \log_{10} E$ as it is more natural to deal with ratios in the log space. We also normalise $\mathbf{N}$ to ensure that the branch lengths $\mathbf{w}$ represent the expected number of mutations. This is done by rescaling $\mathbf{N}$ such that $\sum_{i,i} \pi_i \mathbf{N}_{i,i} = -1$ (Minin et al., 2005).

To simplify the model, we follow Suchard et al. (2003) and impose a product of independent exponential distributions as a prior on the branch lengths $\mathbf{w}$:

$$P(\mathbf{w}|r) = \prod_i P(w_i|r, \boldsymbol{\theta}); \; P(w_i|r) = r^{-1} \exp(-w_i/r), \tag{3}$$

where the w_is are the lengths of the individual branches of the phylogenetic tree. This prior is conjugate to the likelihood and makes analytical integration of the branch lengths tractable:

$$\begin{aligned} P(\mathbf{y}_t|H_S, r, \boldsymbol{\theta}) &= \int P(\mathbf{y}_t|H_S, \mathbf{w}, \boldsymbol{\theta}) \prod_i P(w_i|r)\, d\mathbf{w} \\ &= \pi_{y_{t,1}} \sum_{y_{t,m+1}} \cdots \sum_{y_{t,m+a}} \prod_{(i,j)\in H_S} \mathbf{M}_{y_{t,i},y_{t,j}}(r), \end{aligned} \tag{4}$$

where $y_{t,m+1}$ to $y_{t,m+a}$ represent the nucleotides at position t of the a unknown ancestral sequences, and H_S describes the tree as a list of connections between the sequences defined by the topology of the phylogenetic tree. Suchard et al. (2003) have derived $\mathbf{M}(r) = \sum_{i=1}^{4} (1 + r\lambda_i/\beta)^{-1} \mathbf{u}_i \mathbf{v}_i^T$, where $\mathbf{N} = \sum_{i=1}^{4} \lambda_i \mathbf{u}_i \mathbf{v}_i^T$ is a decomposition of the nucleotide substitution matrix in Equation (2) and $\mathbf{u}_i$, $\mathbf{v}_i$, and λ_i are all functions of $\boldsymbol{\theta}$, the parameters of the nucleotide substitution model – see Hasegawa et al. (1985). We use the pruning algorithm of Felsenstein (1981) to reorder the terms and summations in Equation (4) such that summations involving the smallest number of terms are evaluated first. This reduces the computational cost of evaluating the expression from exponential to polynomial. The hyperparameter r is a scale parameter representing the expected number of mutations over all branches. Note that an uninformative prior is uniform on a log rather than linear scale, and it is therefore more natural to parameterise the model in terms of log r rather than r itself. We therefore define: $\rho_R = \log_{10} r$.

The iid assumption underlying Equation (1) is violated in the presence of recombination, rate variation or changes in the transition-transversion ratio. We first look at modelling topology changes caused by recombination. We assume we know the set of possible tree topologies $\boldsymbol{\rho}_S = \{\rho_{S,1}, \ldots, \rho_{S,k_S}\}$, where k_S is the number of different topologies. For simplicity of implementation, we follow Husmeier (2005) and deal only with 4 sequences in the alignment, so $\boldsymbol{\rho}_S = \{\rho_{S,1}, \rho_{S,2}, \rho_{S,3}\}$ exhaustively covers all possible unrooted tree topologies. A method to select a suitable candidate set $\boldsymbol{\rho}_S$ for alignments with more sequences is suggested in Minin et al. (2005) and straightforward to integrate into our model (although our current software does not support it yet). We introduce the site-dependent discrete hidden state $H_{S,t}$, where $H_{S,t} \in \boldsymbol{\rho}_S$ represents the topology for site t. So, if $H_{S,t} = \rho_{S,i}$, then the topology at alignment position t is $\rho_{S,i}$.

Given k_R log mean branch lengths $\boldsymbol{\rho}_R = \{\rho_{R,1}, \ldots, \rho_{R,k_R}\}$ and k_T log transition-transversion ratios $\boldsymbol{\rho}_T = \{\rho_{T,1}, \ldots, \rho_{T,k_T}\}$, we allow for rate variation and changes in the transition-transversion ratio by associating each alignment position t with hidden random variables $H_{R,t} \in \boldsymbol{\rho}_R$ and $H_{T,t} \in \boldsymbol{\rho}_T$. These pick a rate ρ_R from $\boldsymbol{\rho}_R$ and a log transition-transversion ratio from $\boldsymbol{\rho}_T$ respectively. As before, if $H_{R,t} = \rho_{R,i}$ and $H_{T,t} = \rho_{T,j}$, then at alignment position t the rate is $\rho_{R,i}$ and the log transition transversion-ratio is $\rho_{T,j}$. This allows the mean rate and the transition-transversion ratio to vary along the sequence alignment. To summarise, for a site t in the alignment there are in total three hidden variables: $H_{S,t}$, $H_{R,t}$, and $H_{T,t}$, giving us three *a priori* independent hidden chains. Note that for notational conciseness, we have merged hidden states and their associated parameters into quantities that we refer to as "hidden variables".

In this paper the subscript A will be used to refer to any $A \in \{S, R, T\}$, where S, R, and T refer to states that represent the different tree topologies, rates, and transition-transversion ratios, respectively. We use this notation to define the chains of hidden variables: $\mathbf{H}_A = \{H_{A,1}, \ldots, H_{A,N}\}$, and we represent all hidden variables in the model by defining $\mathbf{h} = \{\mathbf{H}_S, \mathbf{H}_R, \mathbf{H}_T\}$. To allow for correlations between sites that are close together in the sequence – while keeping the computational complexity limited – a Markovian dependence structure is introduced:

$$P(\mathbf{H}_A | k_A, \boldsymbol{\rho}_A) = P(H_{A,1}, \ldots, H_{A,N} | k_A, \boldsymbol{\rho}_A) = \prod_{t=2}^{N} P(H_{A,t} | H_{A,t-1}, k_A, \boldsymbol{\rho}_A) \, P(H_{A,1} | k_A, \boldsymbol{\rho}_A), \tag{5}$$

where again $A \in \{S, R, T\}$. Following Felsenstein and Churchill (1996), the transition probabilities $\boldsymbol{\nu} = \{\nu_S, \nu_R, \nu_T\}$ are defined as:

$$P(H_{A,t} | H_{A,t-1}, \nu_A, k_A, \boldsymbol{\rho}_A) = (\nu_A)^{\mathbb{I}(H_{A,t} = H_{A,t-1})} \left(\frac{1 - \nu_A}{k_A - 1} \right)^{[1 - \mathbb{I}(H_{A,t} = H_{A,t-1})]} \quad \text{for } k_A > 1, \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The parameters ν_S, ν_R, and ν_T denote the probabilities of the tree topology, rate, and transition-transversion ratio, respectively, not changing between adjacent sites. We follow Husmeier and Wright (2001) and set the initial state probabilities to:

$$P(H_{A,1} | k_A) = \frac{1}{k_A}. \tag{7}$$

The resulting model is a FHMM – as illustrated in Figure 1 – containing three *a priori* independent chains of hidden states, $\mathbf{H}_S$, $\mathbf{H}_R$, and $\mathbf{H}_T$, for the tree topologies, evolutionary

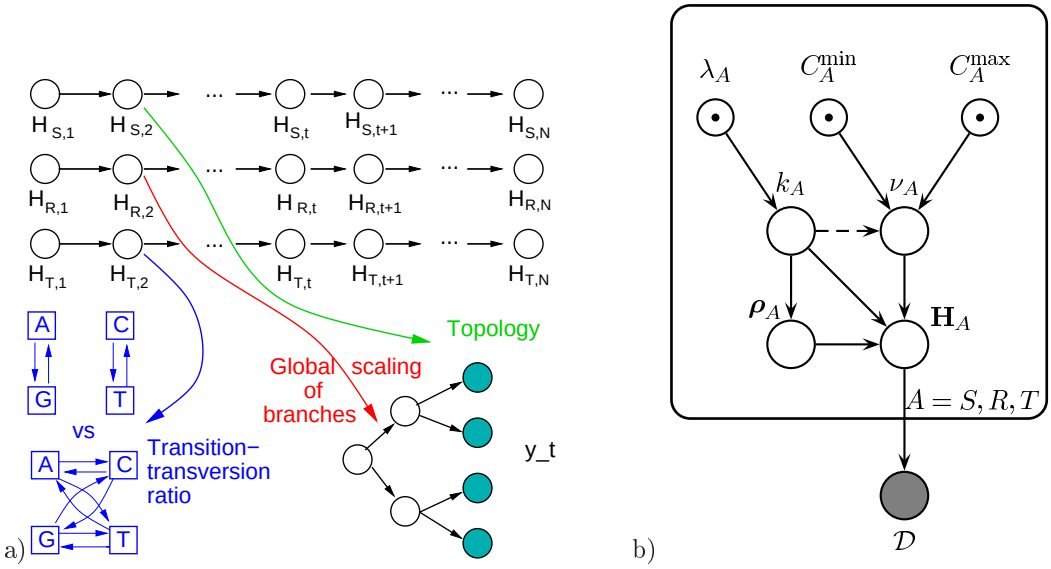

Figure 1. Illustration of the phylogenetic FHMM. Empty circles represent parameters or hidden variables, filled circles indicate observed variables, and dotted circles indicate specified parameters. Sub-figure a) shows how for each position in the alignment, there are three hidden variables representing the topology, the evolutionary rate, and the transition-transversion ratio, and that these hidden variables are correlated between neighbouring positions. These specify the characteristics of the phylogenetic tree, shown in the bottom right, in which empty nodes represent the nucleotides of unobserved ancestral sequences, while shaded nodes represent nucleotides in the DNA sequence alignment. The topology $H_{S,t}$ specifies the connectivity of this tree while the log rate $H_{R,t}$ specifies how likely mutations are along each branch in the tree. The relative likelihood of seeing a transition as opposed to a transversion along each branch is specified by $H_{T,t}$ – the difference is illustrated in the bottom left (squares represent nucleotides). Sub-figure b) shows a representation of our model in the form of a probabilistic graphical model (Pearl, 1988), where $\mathbf{H}_S$, $\mathbf{H}_R$, $\mathbf{H}_T$, and $\mathcal{D}$ are all chains of hidden states, as shown in Sub-figure a). The rounded box is a plate, used to repeat the same nodes three times for $A \in \{S, R, T\}$ – however note that k_S and $\boldsymbol{\rho}_S$ are not inferred by the construction of our model. ν_A is not defined when $k_A = 1$, which we symbolise with a dashed line.

rates, and transition-transversion ratios, respectively. Using FHMMs has two main benefits for our model: efficient sampling from the marginal distribution of $\mathbf{H}_A$, and efficient integration over all possible $\mathbf{H}_A$.

The probability of a column of nucleotides in the alignment, the so-called emission probability, depends on all three hidden states: $P(\mathbf{y}_t|H_{S,t}, H_{R,t}, H_{T,t})$, which can then be calculated using Equation (4). The λ_i's, $\mathbf{u}_i$'s, and $\mathbf{v}_i$'s terms in Equation (4) also depend on the equilibrium nucleotide frequencies (π_A, π_C, π_G, and π_T) – see Hasegawa et al. (1985). However, in order to keep the notation simple, we do not make this dependence explicit in the equations.

Note that, in principle, π_A, π_C, π_G, and π_T from $\boldsymbol{\theta}$ in Equation (4) should be included in our inference scheme, as described in Husmeier and McGuire (2003). However, Husmeier and McGuire (2003) found that in practice a fixation of π_A, π_C, π_G, and π_T at values estimated from their occurrences in the alignment makes little difference to the prediction

of $P(H_{S,t}|\mathcal{D})$, $P(H_{R,t}|\mathcal{D})$, and $P(H_{T,t}|\mathcal{D})$, and has the advantage of reduced computational costs.

2.2. Prior distributions

We introduce prior probabilities on the transition parameters $\boldsymbol{\nu}$: $P(\nu_S)$, $P(\nu_R)$, and $P(\nu_T)$. As shown in Husmeier and McGuire (2003), the conjugate prior is a beta distribution:

$$\mathcal{B}(x;\alpha,\beta) \propto x^{\alpha-1}(1-x)^{\beta-1}, \tag{8}$$

whose shape is determined by the hyperparameters α and β. In the present work, we set $\alpha = \beta = 1$, reducing our prior to a uniform distribution over the interval [0,1], where we additionally constrain the range of valid values:

$$P(\nu_A) \quad \propto \quad \mathcal{B}(\nu_A|\alpha,\beta)\mathbb{I}\left(C_A^{\min} \leq \nu_A \leq C_A^{\max}\right), \tag{9}$$

the reason for this will become clear in Section 6. We define $\mathbf{C} = \left\{C_A^{\min}, C_A^{\max} | A \in \{S,R,T\}\right\}$ to be all such thresholds. ν_A defines a geometric distribution over n_A, the segment length: $P(n_A) = (\nu_A)^{n_A-1}(1-\nu_A)$, which implies an average segment length of $\mathbb{E}[n_A] = 1/(1-\nu_A)$. Thus setting $C_A^{\min}$ and $C_A^{\max}$ implies that the average segment length is between $1/1-C_A^{\min}$ and $1/1-C_A^{\max}$, allowing an intuitive specification of prior knowledge. Also, the posterior distributions of ν_S, ν_R, and ν_T can contain multiple modes, which can be easily selected and more closely investigated by setting $C_A^{\min}$ and $C_A^{\max}$ appropriately.

We set our prior belief on k_R, the number of rate states and k_T, the number of transition-transversion ratios to be:

$$P(k_R) \propto \frac{(\lambda_R)^{(k_R-1)}}{(k_R-1)!}\mathbb{I}(k_R - 1 \leq k_{\max}) \qquad P(k_T) \propto \frac{(\lambda_T)^{(k_T-1)}}{(k_T-1)!}\mathbb{I}(k_T - 1 \leq k_{\max}) \tag{10}$$

which are truncated Poisson distributions over the number of additional rate and transition-transversion ratio states. These are distributions over the number of additional states as the model is nonsensical without at least a single rate and transition-transversion ratio. We expect an average of λ_R additional rate states and λ_T additional transition-transversion states. For instance, if we expect that on average a new rate state occurs every thousand alignment columns, then we could set $\lambda_R = \frac{N}{1000}$. These priors are truncated Poisson distributions, where the truncation ($k_{\max} = 15$ in our case) reflects our desire for a parsimonious solution, with re-use of rates and transition-transversion ratios for different parts of the alignment. In practice, we never observed k_R or k_T to be as high as $k_{\max}$, implying that our model was not affected by this truncation.

To complete the specification of our probabilistic model, we specify priors for $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$, which have k_R and k_T entries respectively:

$$P(\boldsymbol{\rho}_A|k_A) \quad = \quad \prod_{i=1}^{k_A} Q_A(\rho_{A,i}), \tag{11}$$

where Q_A is the distribution over a single entry, for $A \in \{R,T\}$. We set $Q_A(\rho_{A,i}) = \mathcal{N}\left(\rho_{A,i};\mu_A,\sigma_A^2\right)$ where $\mathcal{N}(x;\mu,\sigma^2) \propto \exp\left(-\frac{1}{2}(x-\mu)^2/\sigma^2\right)$ is the Gaussian density function. For comparability, we set these hyperparameters to the values used by Minin et al. (2005): $\mu_R = -2\log_{10} e$, $\sigma_R^2 = 2\log_{10} e$, $\mu_T = 2\log_{10} e$, and $\sigma_T^2 = 1\log_{10} e$. Using uniform

and even-numbered order statistics (Green, 1995) priors on $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$ did not strongly alter our predictions – see Lehrach (2007) for a comparison of these different priors.

In earlier applications of our method we imposed an ordering constraint on the log rates $\boldsymbol{\rho}_R$ for identifiability, as suggested in Green (1995). However, there has been some controversy about the usefulness of imposing such an artificial identifiability constraint Jasra et al. (2005); Celeux et al. (2000). Also, our own investigations have shown, both theoretically as well as empirically, that imposing an ordering constraint has no effect on the quantities of interest estimated with our method (see Lehrach, 2007). We have therefore dropped the ordering constraint altogether, thereby simplifying the method.

In summary, the full prior distribution for the model is:

$$\begin{aligned}\mathcal{P} = P\left(\boldsymbol{\nu}, k_R, k_T, \boldsymbol{\rho}_R, \boldsymbol{\rho}_T, \mathbf{h}, |k_S, \boldsymbol{\rho}_S, \mathbf{C}\right) = P\left(\nu_S | k_S, C_S^{\min}, C_S^{\max}\right) \times \\ \prod_{A \in \{R,T\}} P\left(k_A\right) P\left(\boldsymbol{\rho}_A | k_A\right) P\left(\nu_A | k_A, C_A^{\min}, C_A^{\max}\right) \times \\ \prod_{A \in \{S,R,T\}} \left\{ P\left(H_{A,1} | k_A, \boldsymbol{\rho}_A\right) \prod_{t=2}^{N} P\left(H_{A,t} | H_{A,t-1}, \nu_A, k_A, \boldsymbol{\rho}_A\right) \right\}, \end{aligned} \tag{12}$$

as defined in Equations (6), (7), (9), (10), and (11). We do not define prior distributions over k_S and $\boldsymbol{\rho}_S$ as these parameters are not changed within our model, owing to the fact that the number of different tree topologies is fixed.

2.3. Likelihood

Our likelihood is:

$$\mathcal{L} = P\left(\mathcal{D} | \mathbf{h}\right) = \prod_{t=1}^{N} P\left(\mathbf{y}_t | H_{S,t}, H_{R,t}, H_{T,t}\right), \tag{13}$$

as defined in Equation (4).

2.4. Posterior inference

In the Bayesian paradigm, we are interested in the posterior distribution of the parameters and the hidden variables:

$$P\left(\mathbf{h}, \boldsymbol{\nu}, k_R, k_T, \boldsymbol{\rho}_R, \boldsymbol{\rho}_T | \mathcal{D}, k_S, \boldsymbol{\rho}_S, \mathbf{C}\right) \propto \mathcal{P} \times \mathcal{L}, \tag{14}$$

where the prior $\mathcal{P}$ and likelihood $\mathcal{L}$ are defined in Equations (12) and (13). Recall that for certain bacteria and viruses the tree topology along the alignment can change as a consequence of recombination. This corresponds to a state transition $H_{S,t} = \rho_{S,i} \rightarrow H_{S,t+1} = \rho_{S,(k \neq i)}$ at the breakpoint t of the affected region. Likewise, different segments of a DNA sequence alignment can be under different selective pressure, which corresponds to transitions between different rate states $H_{R,t}$. Hence, our main objective is the estimation of the marginal posterior probabilities:

$$P\left(H_{A,t} | \mathcal{D}, k_S, \boldsymbol{\rho}_S\right) \quad = \quad \sum_{H_{A,1}} \cdots \sum_{H_{A,t-1}} \sum_{H_{A,t+1}} \cdots \sum_{H_{A,N}} P\left(\mathbf{H}_A | \mathcal{D}, k_S, \boldsymbol{\rho}_S\right), \tag{15}$$

where again $A \in \{S, R, T\}$, and the dependence on $\mathbf{C}$ has not been made explicit to simplify the notation. Plotting the distributions of $H_{S,t}$, $H_{R,t}$, and $H_{T,t}$ along the DNA sequence

alignment gives clear indications about the location of recombinant regions, differently diverged regions and regions with changes in the transition-transversion ratio respectively. The distributions $P\left(\mathbf{H}_A|\mathcal{D}, k_S, \boldsymbol{\rho}_S\right)$ are obtained by marginalisation of the posterior:

$$P\left(\mathbf{H}_A|\mathcal{D}, k_S, \boldsymbol{\rho}_S\right) = \sum_{\mathbf{H}_{\{S,R,T\}\setminus A}} \sum_{k_R} \sum_{k_T} \int d\boldsymbol{\rho}_R \int d\boldsymbol{\rho}_T \int d\boldsymbol{\nu} P\left(\mathbf{h}, \boldsymbol{\nu}, k_R, k_T, \boldsymbol{\rho}_R, \boldsymbol{\rho}_T|\mathcal{D}, k_S, \boldsymbol{\rho}_S\right), \tag{16}$$

where we have introduced the notation $\{S, R, T\} \setminus A$ to represent the set of factors excluding the factor A. Hence, when A is S, the above marginalisation reduces to summing over $\mathbf{H}_R$ and $\mathbf{H}_T$. While the marginalisation in Equation (15) can be carried out efficiently with linear time complexity using dynamic programming techniques discussed in Rabiner (1989), the implicit marginalisations to make Equation (14) into a distribution, and the explicit marginalisations in Equation (16) are intractable and have to be numerically approximated with MCMC. For fixed $\boldsymbol{\rho}_R$ and k_R, Husmeier (2005) demonstrated a Gibbs sampling procedure (Casella and George, 1992) for $\mathbf{H}_S$, $\mathbf{H}_R$, ν_S, and ν_R (the transition-transversion ratio was assumed to be invariant along the alignment). However, it is computationally intractable to directly sample from the appropriate marginal distributions of $\boldsymbol{\rho}_R$, k_R, $\boldsymbol{\rho}_T$, or k_T. The novelty of our paper comes from extending Husmeier (2005) to rigorously marginalise over $\boldsymbol{\rho}_R$, $\boldsymbol{\rho}_T$, k_R, and k_T by adopting a RJMCMC Metropolis-Hastings scheme (Green, 1995), allowing us to generate samples for $\boldsymbol{\rho}_R$, k_R, $\boldsymbol{\rho}_T$, and k_T despite the dimensionality of the parameter space changing. The motivation for our model comes from Boys and Henderson (2004), who applied RJMCMC to inference in non-phylogenetic HMMs for segmenting individual DNA sequences. We refer to our model, which generalises this approach to the segmentation of whole DNA sequence alignments in a phylogenetic context, as the Phylogenetic Reversible Jump Factorial Hidden Markov model (PRJ-FHMM).

3. Outline of the MCMC scheme

The following moves are performed for each iteration of the MCMC sampler:

(a) Sample $\mathbf{H}_A \sim P\left(\cdot|\nu_A, \boldsymbol{\rho}_A, k_A, \mathbf{H}_{\{S,R,T\}\setminus A}, \mathcal{D}\right)$ for $A = \{S, R, T\}$.
(b) Sample $\nu_A \sim P\left(\cdot|C_A^{\min}, C_A^{\max}, k_A, \mathbf{H}_A, \mathcal{D}\right)$ for $A = \{S, R, T\}$.
(c) For $A = \{R, T\}$: propose $\boldsymbol{\rho}_A^*$ and k_A^* by adapting $\boldsymbol{\rho}_A$ and k_A. Propose $\mathbf{H}_A^* \sim P\left(\cdot|\nu_A, \boldsymbol{\rho}_A^*, k_A^*, \mathbf{H}_{\{S,R,T\}\setminus A}, \mathcal{D}\right)$. Accept $\boldsymbol{\rho}_A^*$, k_A^* and $\mathbf{H}_A^*$ if $\mathcal{U}[0,1] <$ acceptance probability, where $\mathcal{U}[0,1]$ is a sample from the uniform distribution over the unit interval.

Note that the conditioning part of each distribution contains the Markov blanket (Pearl, 1988) of the respective random variable to be sampled. The Markov blanket is the set of parents, co-parents and children of a node. This set shields off a given node from all the other nodes in the domain, that is, conditional on its Markov blanket, a node is independent of all the other nodes. Hence, conditioning on the Markov blanket is equivalent to conditioning on the complete set of random variables (excluding the variable to be sampled). The Markov blanket of each random variable can easily be read off from Figure 1b: B is in A's Markov blanket if and only if there is either an edge between A and B, or both A and B are parents of another random variable (Pearl, 1988). This proves that the proposed scheme is a valid Gibbs sampling scheme.

3.1. Sampling $\mathbf{H}_A \sim P\left(\cdot|\nu_A, \boldsymbol{\rho}_A, k_A, \mathbf{H}_{\{S,R,T\}\setminus A}, \mathcal{D}\right)$

Sampling the hidden state sequences $\mathbf{H}_S$, $\mathbf{H}_R$, and $\mathbf{H}_T$ can be effected with a Gibbs-within-Gibbs procedure, as described in Husmeier and McGuire (2003). However, the stochastic forward-backward algorithm of Boys et al. (2000) has proven to lead to faster mixing and convergence of the Markov chain (Werhli et al., 2006) and was, thus, used in the simulations reported in this paper.

3.2. Sampling $\nu_A \sim P\left(\cdot|C_A^{\min}, C_A^{\max}, k_A, \boldsymbol{\rho}_A, \mathbf{H}_A, \mathcal{D}\right)$

The sampling steps for ν_S, ν_R, and ν_T are straightforward due to the conjugacy of the beta distribution $\mathcal{B}$, as defined in Equation (8). Define:

$$\Psi_A \;=\; \sum_{t=1}^{N-1} \mathbb{I}\left(H_{A,t} = H_{A,t+1}\right), \qquad \overline{\Psi}_A \;=\; N - 1 - \Psi_A. \tag{17}$$

It is then easy to show from (6) that:

$$P\left(\nu_A|C_A^{\min}, C_A^{\max}, \mathbf{H}_A, k_A, \mathcal{D}\right) \;\propto\; \mathbb{I}\left(C_A^{\min} \leq \nu_A \leq C_A^{\max}\right) \mathcal{B}\left(\nu_A|\Psi_A + \alpha, \overline{\Psi}_A + \beta\right). \tag{18}$$

See Husmeier and McGuire (2003) for a derivation for the untruncated case. If $k_A = 1$, then Equation (18) does not apply, as there is only a single possible $\mathbf{H}_A$. To generate a sample from the truncated beta distribution we use a simple Metropolis-Hastings method – see Lehrach (2007) for more detail.

Additionally, Ψ_A and $\overline{\Psi}_A$ allow us to generate an estimate of the posterior distribution of ν_A:

$$P\left(\nu_A|\mathcal{D}, C_A^{\min}, C_A^{\max}\right) \approx \frac{1}{M}\sum_{i=1}^{M} \frac{\mathcal{B}\left(\nu_A|\Psi_A^{(i)} + \alpha, \overline{\Psi}_A^{(i)} + \beta\right)}{\int_{C_A^{\min}}^{C_A^{\max}} \mathcal{B}\left(\nu_A|\Psi_A^{(i)} + \alpha, \overline{\Psi}_A^{(i)} + \beta\right) d\nu_A} \mathbb{I}\left(C_A^{\min} \leq \nu_A \leq C_A^{\max}\right), \tag{19}$$

where the superscript (i) represents the i^{th} sample and we have M samples. The integral is easily calculated using the trapezoid method.

3.3. Proposing and conditionally accepting $\boldsymbol{\rho}_A^*$, k_A^*, and $\mathbf{H}_A^*$

We adopt a Reversible Jump Metropolis-Hastings scheme (Green, 1995) where we propose a new number of rate states $k_R{}^*$ and a new set of rate states $\boldsymbol{\rho}_R{}^*$ from k_R and $\boldsymbol{\rho}_R$. This is done using a birth move (with probability b_k), a death move (with probability d_k) or a relocation of one of the rate states (with probability r_k). A new $\mathbf{H}_R{}^*$ is then proposed given the new $\boldsymbol{\rho}_R{}^*$. The new set of rate states $\boldsymbol{\rho}_R{}^*$ is then accepted with a probability such that given ergodicity, the Markov chain is guaranteed to converge in distribution to the correct posterior distribution. This procedure is similar to the reversible jump move (b) from Boys and Henderson (2004).

$\boldsymbol{\rho}_T$ and k_T are adapted in the same way with identical derivations, so we only show the derivation for $\boldsymbol{\rho}_R$ and drop the R subscript on k_R. To use the Reversible Jump method, we need to specify how we propose k^* and $\boldsymbol{\rho}_R^*$. The set of all possible proposal moves is outlined in Table 1. Note that k^* is proposed such that the Hastings factor cancels out against the prior ratio. Lastly, we propose $\mathbf{H}_R^* \sim P\left(\cdot|\nu_R, \boldsymbol{\rho}_R^*, k^*, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D}\right)$ as described in Section 3.1.

Table 1. Possible proposal moves, the probability with which they are selected, and the corresponding proposal probability $\pi_M\left(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R\right)$ for $\boldsymbol{\rho}_R{}^*$. All π distributions presume that $\boldsymbol{\rho}_R{}^*$ is a valid proposal given the move type, as otherwise the π distributions are not normalised. We use $c = 0.4$ – see Green (1995).

Move type	Probability of move and proposal for $\boldsymbol{\rho}_R{}^*$	Description of how $\boldsymbol{\rho}_R{}^*$ is proposed
Birth $k^* = k+1$	$b_k = c\min\left\{1, \frac{P(k+1)}{P(k)}\right\}$ $\pi_b\left(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R\right) = \frac{1}{k+1}Q\left(\rho_R{}^*\right)$	A new rate is sampled from Q in Equation (11), the prior distribution on $\boldsymbol{\rho}_R$ for a single rate. Where to insert the new rate state is randomly and uniformly sampled from the $k+1$ possibilities.
Death $k^* = k-1$	$d_k = c\min\left\{1, \frac{P(k-1)}{P(k)}\right\}$ $\pi_d\left(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R\right) = \frac{1}{k}$	A randomly chosen rate is deleted.
Relocation $k^* = k$	$r_k = 1 - (b_k + d_k)$ $\pi_r\left(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R\right) = \frac{1}{k}Q\left(\rho_R{}^*\right)$	An existing rate factor position is randomly chosen, and its position re-sampled from Q (see birth move).

The acceptance probability a of k^*, $\boldsymbol{\rho}_R{}^*$, and $\mathbf{H}_R{}^*$ is $\min\{1, A_B\}$, where:

$$A_B = \text{Likelihood ratio} \times \text{Prior ratio} \times \text{Inverse proposal probability ratio} \times \left|\det\left(\text{Jacobian}\right)\right|, \tag{20}$$

see Green (1995) – our formulation is closer to that of Suchard et al. (2003). We first derive the acceptance probability of a birth move. We first propose a new rate state $\rho_R{}^*$ from Q_R in Equation (11). We then map $(\boldsymbol{\rho}_R, \rho_R{}^*)$ to $(\boldsymbol{\rho}_R{}^*)$. In Equation (20), the Jacobian term refers to this mapping, and det stands for the determinant. This mapping is a permutation, hence the Jacobian is a permutation matrix, which implies $\det\left(\text{Jacobian}\right) = \pm 1$, so $\left|\det\left(\text{Jacobian}\right)\right| = 1$.

From Equations (11), (12), and (13), and Table 1 we see that after cancelling, the terms (by their initials) are:

$$\begin{aligned}
\text{LR} &= \frac{P\left(\mathcal{D}|\mathbf{H}_R, k, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T\right)}{P\left(\mathcal{D}|\mathbf{H}_R{}^*, k+1, \nu_R, \boldsymbol{\rho}_R^*, \mathbf{H}_S, \mathbf{H}_T\right)} \\
\text{PR} &= \frac{P\left(\mathbf{H}_R|k, \nu_R, \boldsymbol{\rho}_R\right)}{P\left(\mathbf{H}_R{}^*|k, \nu_R, \boldsymbol{\rho}_R\right)} \frac{P\left(k+1\right)}{P\left(k\right)} \frac{P\left(\boldsymbol{\rho}_R{}^*|k+1\right)}{P\left(\boldsymbol{\rho}_R|k\right)} \\
&= \frac{P\left(\mathbf{H}_R|k, \nu_R, \boldsymbol{\rho}_R\right)}{P\left(\mathbf{H}_R{}^*|k, \nu_R, \boldsymbol{\rho}_R\right)} \frac{P\left(k+1\right)}{P\left(k\right)} \left[Q\left(\rho_R{}^*\right)\right] \\
\text{IPPR} &= \frac{P\left(\mathbf{H}_R|k, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D}\right)}{P\left(\mathbf{H}_R{}^*|k+1, \nu_R, \boldsymbol{\rho}_R^*, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D}\right)} \frac{d_{k+1}\pi_d\left(\boldsymbol{\rho}_R|\boldsymbol{\rho}_R{}^*\right)}{b_k\pi_b\left(\boldsymbol{\rho}_R{}^*|\boldsymbol{\rho}_R\right)} \\
&= \frac{P\left(\mathbf{H}_R|k, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D}\right)}{P\left(\mathbf{H}_R{}^*|k+1, \nu_R, \boldsymbol{\rho}_R^*, \mathbf{H}_S, \mathbf{H}_T, \mathcal{D}\right)} \frac{P\left(k\right)}{P\left(k+1\right)} \frac{k+1}{k+1} \frac{1}{Q\left(\rho_R{}^*\right)}.
\end{aligned}$$

All terms not involving $\mathcal{D}$ and $\mathbf{H}_R$ cancel between PR and IPPR. LR and PR together form a ratio of joint distributions over $\mathcal{D}$ and $\mathbf{H}_R$ which in turn simplifies against the ratio of distributions over $\mathbf{H}_R$ conditioned on $\mathcal{D}$ in IPPR. Hence:

$$A_B = \frac{P\left(\mathcal{D}|k+1, \nu_R, \boldsymbol{\rho}_R{}^*, \mathbf{H}_S, \mathbf{H}_T\right)}{P\left(\mathcal{D}|k, \nu_R, \boldsymbol{\rho}_R, \mathbf{H}_S, \mathbf{H}_T\right)}, \tag{21}$$

where due to the HMM structure, $P\left(\mathcal{D}|\boldsymbol{\rho}_R{}^*, k+1, \mathbf{H}_S, \mathbf{H}_T, \nu_R\right)$ can be computed from Equations (12) and (13) in linear time with a dynamical programming algorithm known as

the forward algorithm (Rabiner, 1989). Note that the stated dependence on the conditioning variables becomes clear from the conditional independence graph of Figure 1b and the properties of the Markov blanket, as discussed above.

The same cancellations and simplifications occur when considering the acceptance probability of the death move as the death move is the inverse of the birth move. Hence the acceptance probability of a death move is the same (after replacing $k^* = k + 1$ with $k^* = k - 1$) . The acceptance probability of a relocation is also the same (after replacing $k^* = k + 1$ with $k^* = k$) as relocation moves are symmetrical to themselves, in the same way as the birth and death moves are symmetrical.

Note that Equation (18) does not apply when $k = 1$ as there is only a single possible $\mathbf{H}_R$. Hence ν_R has no effect on the likelihood. When moving from two rates to a single rate state, ν_R is removed from the system. Correspondingly, when moving from a single rate state to two rate states, ν_R is proposed from the prior. To see that this leaves the acceptance ratios unchanged, first consider the death move from two rate states to a single rate. We have an extra $P\left(\nu_R|C_R^{\min}, C_R^{\max}\right)$ in the denominator of PR, and a new proposal term $Q\left(\nu_R\right)$ in the numerator of the IPPR. We set $Q\left(\nu_R\right) = P\left(\nu_R|C_R^{\min}, C_R^{\max}\right)$ so that these terms cancel, leaving the acceptance probability unchanged. The reverse argument applies to the birth move, so the acceptance probability is again unchanged.

3.4. Specific Markov chain settings and convergence diagnostics

To check for convergence, we used the method of Gelman and Rubin (1992) and computed the Potential Scale Reduction Factors (PSRF) of $H_{R,t}$ and $H_{T,t}$ for $t \in \{1, \ldots, N\}$, and ν_A. These characteristics were chosen as they are invariant to the dimensionality of the parameter space. All results presented in the paper, for all models, were run at least in triplicate (with the exception of the initial ν explorations and the synthetic codon effect study, which were repeated 10 times). For our proposed model, the initial number of rates was picked uniformly between 1 and $k_{\max}$, with each rate sampled randomly from the uniform distribution. In this paper, we are mainly interested in investigating the rate along the alignment, so all runs were started with only a single transition-transversion ratio, randomly sampled from the uniform distribution.

We discarded the first 10,000 iterations of the PRJ-FHMM samples as the burn-in period. Then, for the next 200,000 iterations every 10^{th} sample was kept so that we could form the posterior summaries. The MCP of Minin et al. (2005) was run for 200,000 burn-in iterations, followed by 4,000,000 sampling iterations where every 200th sample was kept. These lengths were chosen as they resulted in similar convergence indications, as measured by the PSRF. The high PSRF was consistently less than 1.08 (and often much less), indicating a sufficient degree of convergence. The proposed sampling scheme was extensively tested on synthetic data and its results compared to using numerical integration on simple cases. In order to keep this article concise, these results are presented elsewhere – see Lehrach (2007).

4. Comparison with a breakpoint model

Suchard et al. (2003) introduced the multiple-change point (MCP) model where a DNA sequence alignment is split into discrete segments by a series of breakpoints. The number of segments is thus always one greater than the number of breakpoints. The topology, rate and transition-transversion ratio are estimated independently between each successive pair of breakpoints. This joint estimation causes *a priori* correlation between the locations of the

changes in the rate and the locations of changes in the topology, which Minin et al. (2005) removed by using one MCP to model the topology changes, and a separate MCP to model changes in both the rate and transition-transversion ratio. A software implementation of their model is available from `http://www.biomath.ucla.edu/msuchard/DualBrothers/`.

Minin et al. (2005) place truncated Poisson distributions over b_R, the number of breakpoints of the rate or transition-transversion ratio along the alignment, and b_S, the number of breakpoints of the topology along the alignment:

$$P\left(b_R|\lambda_R^B\right) \propto \frac{\left(\lambda_R^B\right)^{b_R}}{b_R!}\mathbb{I}\left(b_R < N\right), \qquad P\left(b_S|\lambda_S^B\right) \propto \frac{\left(\lambda_S^B\right)^{b_S}}{b_S!}\mathbb{I}\left(b_S < N\right). \tag{22}$$

where λ_R^B and λ_S^B define the *a priori* expected mean numbers of rate/transition-transversion ratio and topology breakpoints respectively.

While the MCP of Minin et al. (2005) separates out estimating the phylogenetic topology, their model still jointly estimates the rate and transition-transversion ratio. The PRJ-FHMM estimates these quantities independently, complicating our theoretical comparison. For the purposes of comparing the models, we will henceforth assume that $\boldsymbol{\rho}_R$ in the PRJ-FHMM has been augmented to additionally contain the transition-transversion ratio associated with each rate, allowing us to ignore ν_T. This simplification of the PRJ-FHMM allows us to make the following observations. In the PRJ-FHMM, ν_R defines a binomial distribution over the number of rate breakpoints in the alignment:

$$P\left(b_R|\nu_R\right) = \left(\begin{array}{c} N-1 \\ b_R \end{array}\right)\left(1-\nu_R\right)^{b_R}\nu_R^{(N-1)-b_R}. \tag{23}$$

where this argument also applies to ν_S (ν_T is assumed to have been merged into ν_R). In the limit of $N \to \infty$ and $\nu_R \to 1$, the distribution in Equation (23) tends to that of (22), with:

$$\nu_R = 1 - \frac{\lambda_R^B}{N-1}, \tag{24}$$

as the Poisson distribution is the limiting case of the binomial distribution. In practice, these distributions are extremely similar for small values of λ_R^B. Hence, the Poisson priors used in the MCPs of Suchard et al. (2003) and Minin et al. (2005) can be regarded as almost equivalent to a single setting of ν_S and ν_R, the transition probabilities in the PRJ-FHMM. To summarise: the MCP of Minin et al. (2005) is effectively a special case of our model with a fixed value of ν_S and ν_R, each state occurring only once and no separation between the processes leading to changed rates and changed nucleotide substitution parameters. In contrast, our PRJ-FHMM separates k_R, the number of states (or different segment types) from the average segment length (determined by ν_R), where the average segment length is not set to a fixed value, but also inferred.

For an empirical comparison between the proposed phylogenetic FHMM and the dual MCP of Minin et al. (2005), we map $\lambda_R^B = 2\lambda_R - 1$. This assumes that on average every extra rate state causes two extra breakpoints and that when $\lambda_R^B = 1$, both models are set to their most conservative prior distributions. We have carried out an extensive comparison between the proposed phylogenetic FHMM and the MCP of Minin et al. (2005). In order to keep the present article concise, we include only a subset of the results here, and refer the reader to Lehrach (2007) for further details.

5. Investigating the coupling of the rate and transition-transversion ratios

We also consider an alternative model where the rate and transition-transversion ratio are coupled. Instead of $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$, we have k_J joint states: $\boldsymbol{\rho}_J = \{\rho_{J,1}, \ldots, \rho_{J,k_J}\}$, where $\rho_{J,i} = [\rho_{R,i}, \rho_{T,i}]$. We now allow for rate variation and changes in the transition-transversion ratio by associating each alignment position t with hidden random variables $H_{J,t} \in \boldsymbol{\rho}_J$. Hence, given that $H_{J,t} = \rho_{J,i}$, where $\rho_{J,i} = [\rho_{R,i}, \rho_{T,i}]$, the rate and the transition-transversion ratio at alignment column t are then $\rho_{R,i}$ and $\rho_{T,i}$. The new priors are:

$$P(k_J) \propto \frac{(\lambda_J)^{k_J-1}}{(k_J-1)!}\mathbb{I}(k_J - 1 \leq k_{\max}), \tag{25}$$

$$P(\boldsymbol{\rho}_J) = \prod_{i=1}^{k_J} Q_R(\rho_{R,i})\, Q_T(\rho_{T,i}) \text{ where } \rho_{J,i} = [\rho_{R,i}, \rho_{T,i}], \tag{26}$$

where Q_R and Q_T are defined in Section 2.2. The prior on k_J is of the same form as the prior on k_R given in Equation 10, while the prior on $\boldsymbol{\rho}_J$ is the product of the priors on the individual terms in $\boldsymbol{\rho}_R$ and $\boldsymbol{\rho}_T$.

The modelling of the topology is unchanged. While this model is also implemented in our software, the focus in this paper is upon the original model. We can easily compute the Bayes factor between these two models. The marginal likelihood of each hypothesis ($\mathcal{H}_1$ = decoupled rate and transition-transversion ratio, $\mathcal{H}_2$ = coupled) can be expressed as:

$$P(\mathcal{D}|\lambda_R, \lambda_T, \mathcal{H}_1) = p_{11} + p_{12} \qquad P(\mathcal{D}|\lambda_J, \mathcal{H}_2) = p_{21} + p_{22}, \tag{27}$$

where:

$$\begin{aligned}
p_{11} &= \int d_{\rho_{R,1}} \int d_{\rho_{T,1}} P(\mathcal{D}, k_R = 1, k_T = 1, \boldsymbol{\rho}_R = \{\rho_{R,1}\}, \boldsymbol{\rho}_T = \{\rho_{T,1}\} | \lambda_R, \lambda_T, \mathcal{H}_1) \\
p_{21} &= \int d_{\rho_{J,1}} P(\mathcal{D}, k_J = 1, \boldsymbol{\rho}_J = \{\rho_{J,1}\} | \lambda_J, \mathcal{H}_2) \\
p_{12} &= \sum_{k_R} \sum_{k_T} \mathbb{I}(k_R > 1 \vee k_T > 1) \int d_{\boldsymbol{\rho}_R} \int d_{\boldsymbol{\rho}_T} P(\mathcal{D}, k_R, k_T, \boldsymbol{\rho}_R, \boldsymbol{\rho}_T | \lambda_R, \lambda_T, \mathcal{H}_1) \\
p_{22} &= \sum_{k_J > 1} \int d_{\boldsymbol{\rho}_J} P(\mathcal{D}, k_J, \boldsymbol{\rho}_J | \lambda_J, \mathcal{H}_2).
\end{aligned}$$

In order to keep the exposition more concise, we have marginalised over $\mathbf{H}_S$, integrated over ν_S, and removed the dependencies on k_S, $\boldsymbol{\rho}_S$, and $\mathbf{C}$ from the notation. We can easily estimate p_{12}/p_{11} and p_{22}/p_{21} from our MCMC simulations as the fraction of samples which satisfy $k_R > 1 \vee k_T > 1$ and $k_J > 1$ respectively. As $P(\mathcal{D}|k_R = 1, k_T = 1, \mathcal{H}_1) = P(\mathcal{D}|k_J = 1, \mathcal{H}_2)$, it is easy to compute p_{11}/p_{21} from Equation (10). For instance, if $\lambda_R = 1$, $\lambda_T = 1$, $\lambda_J = 1$, and $k_{\max} \to \infty$, then $p_{11}/p_{21} = e^{-1}$. We can then compute the Bayes factor between the two models of interest as:

$$\frac{P(\mathcal{D}|\lambda_R, \lambda_T, \mathcal{H}_1)}{P(\mathcal{D}|\lambda_J, \mathcal{H}_2)} = \frac{p_{11} + p_{12}}{p_{21} + p_{22}} = \frac{\frac{p_{11}}{p_{21}} + \frac{p_{12}}{p_{11}}\frac{p_{11}}{p_{21}}}{1 + \frac{p_{22}}{p_{21}}} = \frac{\frac{p_{11}}{p_{21}}\left(1 + \frac{p_{12}}{p_{11}}\right)}{1 + \frac{p_{22}}{p_{21}}}. \tag{28}$$

Additionally, we can also compute the marginal likelihood of each model. This requires computing p_{11} and p_{21}, which can be requires marginalising over $\boldsymbol{\rho}_R$, $\boldsymbol{\rho}_T$, ν_S, and $\mathbf{H}_S$. $\mathbf{H}_S$ can be marginalised over in a computationally efficient manner using the forwards-backwards algorithm, and numerical integration over the remaining 3 variables is a relatively inexpensive computational operation. Given p_{11} and p_{21}, the marginal likelihoods of the models can be computed as $p_{11}\left(1+{}^{p_{12}}/_{p_{11}}\right)$ and $p_{21}\left(1+{}^{p_{22}}/_{p_{21}}\right)$. Note that this numerical integration is not required for computing the Bayes factor between the models.

6. Segmentation of a bacterial DNA sequence alignment

One of the first indications for interspecific recombination was found in the bacterial genus *Neisseria* (Maynard Smith, 1992). We choose a 787 nucleotide subset (between positions 296-1028) of the *Neisseria argF* DNA multiple alignment studied by Zhou and Spratt (1992). We selected the four strains *Neisseria gonorrhoeae* (X64860), *Neisseria meningitidis* (X64866), *Neisseria cinera* (X64869), and *Neisseria mucosa* (X64873), where GenBank/EML accession numbers are shown in brackets.

We investigate the stability of the methods by investigating the settings: $\lambda_R \in \{1,\ldots,5\}$, which we map to $\lambda_R^B \in \{1,3,5,7,9\}$ for the MCP of Minin et al. (2005), as discussed in Section 4. For the MCP, we set $\lambda_S^B = 0.693$ in Equation (22) as suggested by the authors. For the PRJ-FHMM we set $\lambda_T = 1$, as we are mainly interested in the prediction of rate variation.

6.1. The posterior distributions of ν_S and ν_R

If an alignment is best modelled by a single rate state instead of multiple rate states, then changing ν_R has no effect on the system, and is irrelevant. Hence, the posterior probability of ν_A (where $A \in \{S,R,T\}$) being relevant is $\frac{1}{M}\sum_i \mathbb{I}\left(k_A^{(i)} > 1\right)$, where $A \in \{S,R,T\}$ and M is the number of samples. Bearing this in mind, Figure 2 shows the posterior distributions of ν_S and ν_R. ν_R is relevant with probability 0.998 ± 0.003. Hence, it is highly likely that the alignment exhibits rate heterogeneity. ν_S is always relevant, as $\boldsymbol{\rho}_S$ and k_S are fixed by construction of our model.

Notice that the posterior distribution of ν_S has only a single mode, while the posterior distribution of ν_R has multiple modes. The multi-modality is presumably due to a codon position specific rate variation. When a DNA sequence codes for a protein, each triplet in the sequence codes for a single amino acid. A change in the third position of the triplet often does not change which amino acid is coded for. Hence, a mutation can occur in this position without having an impact on the function of the protein. Consequently, nucleotide substitutions in this position get established with a higher probability in the population, resulting in the codon position specific rate variation. To test this conjecture, we synthetically generated a set of DNA sequence alignments with different trade-offs between codon-specific and region-specific rate variation. See Lehrach (2007) for details of how the simulations were carried out. The results are shown in the bottom panels of Figure 2 and suggest that the bimodality observed in the distribution of $P(\nu_R)$ on Neisseria could, indeed, result from an interplay of these two effects. This implies that the peak around $\nu_R = 0.5$ could be due to the codon effect. We are mainly interested in region-specific rate heterogeneity, owing to its confounding effect on the detection of recombination (Husmeier, 2005), and its increasing relevance in functional genomics (Nimrod et al., 2005; Siepel and

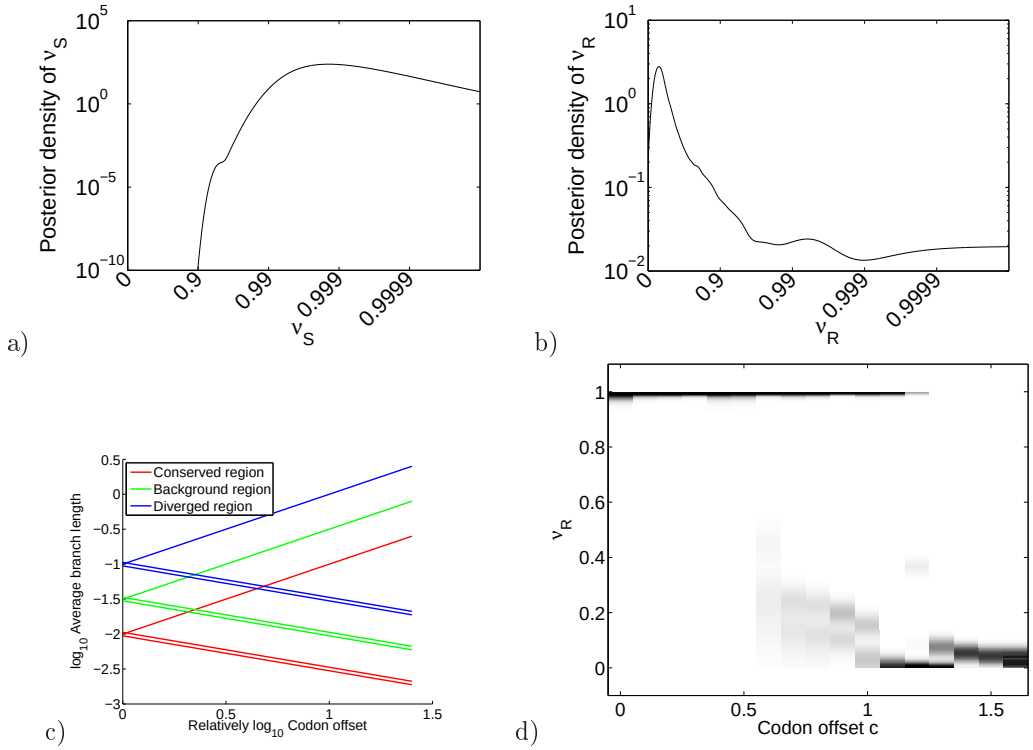

Figure 2. The posterior distributions of ν_S and ν_R for Neisseria, and a synthetic study into the discovered codon effect. In Sub-figure a), ν_S is plotted logarithmically approaching 1 along the x-axis and $P(\nu_S|\mathcal{D})$ is plotted on a logarithmic scale on the y-axis. Sub-figure b) displays $P(\nu_R|\mathcal{D})$, in the same manner as Sub-figure a). The posterior probability of ν_R being relevant for modelling the alignment is 0.998 ± 0.003, where ν_R is irrelevant if there is only a single rate state ($k_R = 1$). Sub-figures c) and d): A synthetic study of a possible reason for the multi-modality of $P(\nu_R)$ seen in Sub-figure b). The log rate of each codon triplet along the sequence is: $\left[\rho_R - \frac{c}{2}, \rho_R - \frac{c}{2}, \rho_R + c\right]$, where ρ_R is the rate of the segment and c reflects the strength of the codon specific behaviour. Sub-figure c) shows the setup of the synthetic study with different shadings indicating different segments. The three lines for each segment indicate the three codon positions. Sub-figure d) shows how the posterior for ν_R varies as we increase c. The horizontal axis represents c while the y-axis displays the posterior distribution of ν_R for that value of c. Darker shadings indicate higher probability. As the codon effect becomes stronger, it starts to dominate the predictions. At $c \approx 0.6$, we can see a multi-modal posterior as the model picks up both behaviours, corresponding to the multiple peaks found in Sub-figure b). This suggests that the peak at around $\nu_R = 0.5$ in Sub-figure b) results from a codon position specific rate variation. Applications in functional genomics are interested in the large-scale effect of rate variation to identify genomic regions under varying degrees of selective pressure. For this reason we set $C_R^{\min}$, the threshold on ν_R, to the minimum of $P(\nu_R|\mathcal{D})$ in Sub-figure b). In this way, we switch off the codon effect, that is, the uninteresting contributions stemming from the signature of the genetic code.

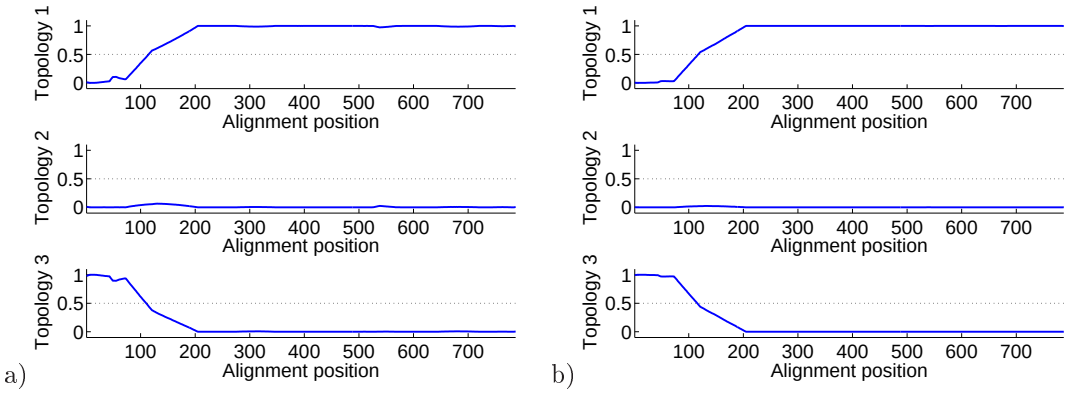

Figure 3. The posterior distribution of the phylogenetic tree topology along the Neisseria alignment. Sub-figure a) shows the resulting estimates for the PRJ-FHMM while Sub-figure b) shows the resulting estimates for the MCP of Minin et al. (2005). The x-axis represents the alignment position, the y-axis the probability, and each sub-plot indicates the posterior probability of each possible topology. Topology 1 : [(*N.gonorrhoeae*, *N.meningitidis*), (*N.cinera*, *N.mucosa*)]; Topology 2 : [(*N.gonorrhoeae*, *N.cinera*), (*N.meningitidis*, *N.mucosa*)]; Topology 3: [(*N.gonorrhoeae*, *N.mucosa*), (*N.cinera*, *N.meningitidis*)]. Zhou and Spratt (1992) predicted a breakpoint at position 202, while the phylogenetic FHMM predicts it to lie in the region 80-200 - this is the region where the posterior probability of the recombinant tree topology gradually decreases from 1 to 0.

Haussler, 2004). For that reason we focus on the peak representing promising long scale behaviour at around $\nu_R = 0.995$.

In contrast to ν_R, the posterior probability of ν_T being relevant is 0.677 ± 0.006, indicating that the model is uncertain about whether or not the transition-transversion ratio changes along the alignment. We do not show the posterior distribution for ν_T as the main focus of this paper is investigating rate variation – see Lehrach (2007) for a more complete investigation of the behaviour of the transition-transversion ratio.

In order to accurately determine the minimum between the mode reflecting the codon position specific rate variation and mode reflecting the long scale behaviour, we reran the simulation with $C_R^{\min} = C_T^{\min} = 0.95$. The resulting posterior distributions over ν_R and ν_T were then used to determine the locations of the minima between the codon and region effect peaks. The lower bounds on ν_R and ν_T were then set to these minima: $C_R^{\min} = 0.975$ and $C_T^{\min} = 0.992$.

6.2. Posterior distribution of the rate and phylogenetic tree topology

In Figure 3, we investigate the posterior distribution of the phylogenetic tree topology for the PRJ-FHMM and the MCP of Minin et al. (2005). The predictions are in good agreement with those of Zhou and Spratt (1992) – see the caption of the figure for details. We display the predictions for $\lambda_R = 1$ and $\lambda_R^B = 1$ – the predictions appeared stable for the ranges of priors on the rates we tested: $\lambda_R \in \{1, \ldots, 5\}$ and the corresponding $\lambda_R^B \in \{1, 3, 5, 7, 9\}$.

Figure 4 compares the estimated rate along the alignment for the PRJ-FHMM and the MCP of Minin et al. (2005). We display the results for the extremities of the prior range. The PRJ-FHMM consistently finds that regions 1-75 and 425-530 are more diverged, with some minor divergence around 345-385. In contrast, the MCP is strongly dependent on

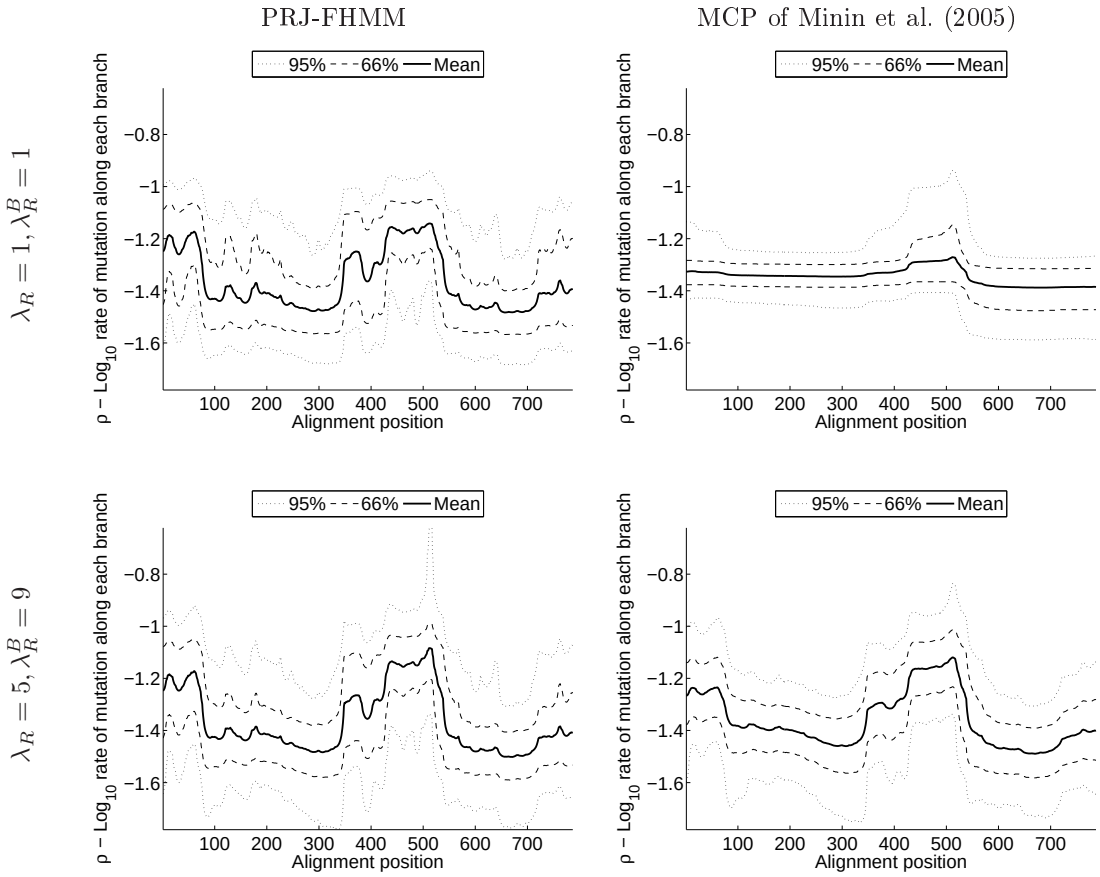

Figure 4. The credibility intervals of the posterior log rate distribution along the Neisseria alignment for the PRJ-FHMM and the MCP of Minin et al. (2005). In each panel, the x-axis indicates the alignment position and the y-axis indicates the log rate. The black line represents the mean posterior rate, while the dashed and dotted lines represent the 66% and 95% credibility intervals. The panels on the left represent the PRJ-FHMM, while the panels on the right represent the MCP. The top and bottom panels compare the lowest and highest values of λ_R (hence also λ_R^B, due to the mapping described in Section 4) that were investigated.

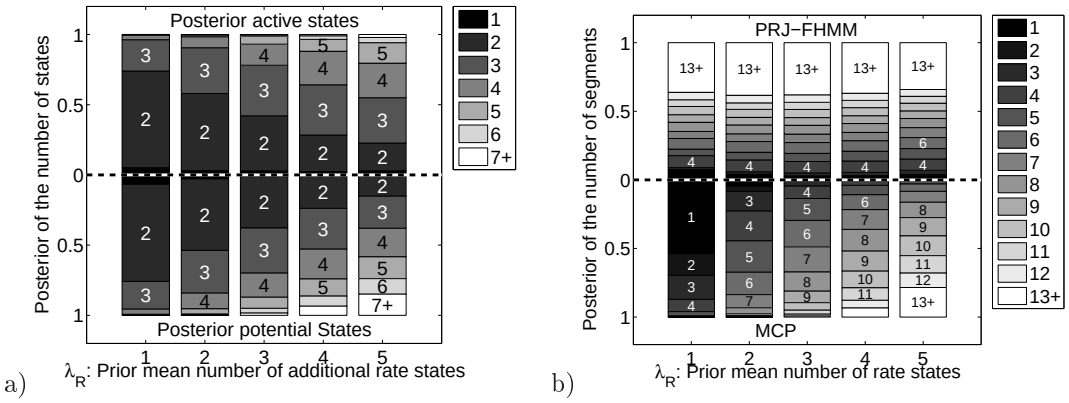

Figure 5. Comparisons of the predicted numbers of rate states and segments present on the alignment of bacterial DNA sequences by our PRJ-FHMM and the MCP of Minin et al. (2005). In both cases, the horizontal axis represents λ_R, the expected mean number of additional rate states, mapped to the MCP as described in Section 4. Sub-figure a) shows the estimated number of states, with active rate states depicted above the line, and potential rate states shown below the line – active rate states must occur at least once along the alignment. For $\lambda_R \in \{1, \ldots, 5\}$, the PRJ-FHMM estimates that there are two to three rate states that occur repeatedly along the alignment. Sub-figure b) shows the estimated number of segments with PRJ-FHMM and MCP (depicted in the top and bottom panel, respectively). The MCP is sufficiently sensitive to changes in λ_R that the estimation of the number of segments present is highly dependent on prior knowledge. The estimation of the distribution over the number of segments for the PRJ-FHMM is stable, with 1, 4 or 6 segments most likely.

the prior as the detection of rate variation at positions 1-75 is found only for specific settings of the prior. Zhou and Spratt (1992) found two anomalous regions present in the alignment: 1-202 and 507-538. They suggested that 1-202 is the result of recombination, while unsure of the origin of the anomalous region 507-538. Both the PRJ-FHMM and the MCP agree with Zhou and Spratt (1992) that no topology change occurs around the region 507-538, but identify the anomalous region as part of a larger region of rate variation, namely 425-530. The PRJ-FHMM and MCP consistently identified that a recombination event occurs towards the beginning of the sequence. However, only the PRJ-FHMM consistently identified the region 1-75 as being more diverged. In contrast, the MCP predictions are dependent on the setting of λ_R^B.

Figure 5 shows comparisons of the predicted numbers of rate states and segments by the PRJ-FHMM and the MCP of Minin et al. (2005), and their dependence on the prior. Both the number of potential and active rate states are shown, where an active rate state is defined as having to occur at least once along the alignment. The estimated number of active rate states is more stable to changes in λ_R. The PRJ-FHMM gives a stable prediction of the number of different rate states present, which neither the MCP nor the model of Husmeier (2005) can estimate. For $\lambda_R \in \{1, \ldots, 5\}$, the PRJ-FHMM predicts that there are two to three rate states that occur repeatedly along the alignment. Sub-figure b) shows the predicted number of segments with the PRJ-FHMM and MCP shown in the top and bottom panels, respectively. The MCP cannot predict the number of states present, and furthermore is sufficiently sensitive to changes in λ_R that predicting the number of

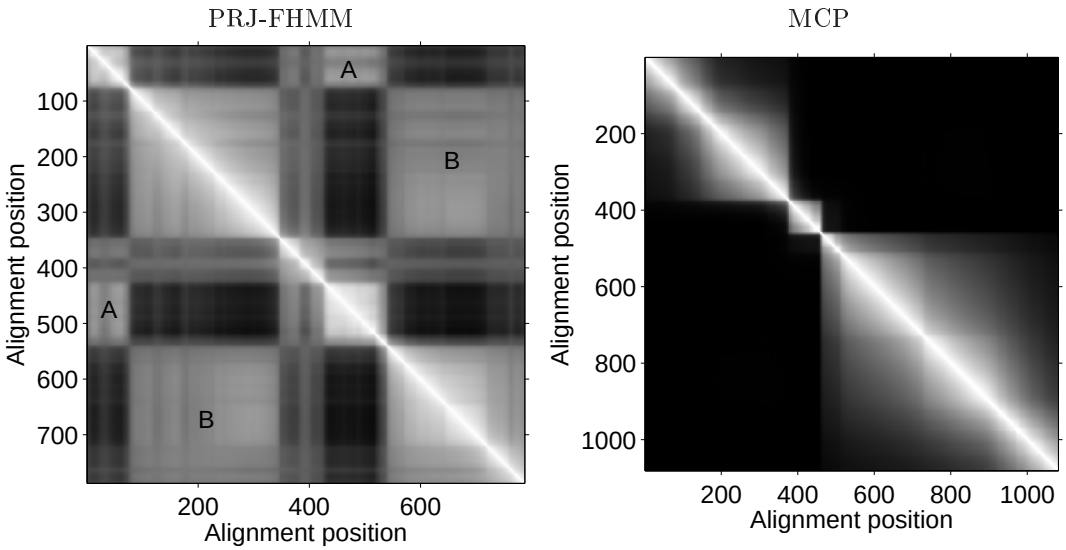

Figure 6. Posterior probability that two alignment positions in Neisseria are in the same rate state. Brighter shading indicates that it is more likely that the alignment positions shown along the x and y axes are in the same rate state.

segments present is extremely difficult due to our lack of knowledge about how λ_R is set. In contrast, the predictions of the PRJ-FHMM are stable – a consequence of the fact that in the PRJ-FHMM, we infer the distribution over ν_R from the DNA sequence alignment. This produces stable predictions which indicate uncertainty over the true number of segments present.

6.3. Posterior probability of alignment positions being in the same state.

Figure 6 shows the posterior probability of pairings of columns in the alignment sharing the same rate state. When we set $\lambda_R^B = 1$, the MCP did not pick up on rate variation around positions 1-75. Hence we instead show the results for $\lambda_R^B = 9$ and $\lambda_R = 5$. These plots are symmetrical, and the posteriors along the diagonal are all 1 as they are by definition in the same state as themselves.

The diagonal squares capture the trivial effect that the evolutionary histories and nucleotide distributions of consecutive sites tend to be similar. The interesting information is contained in the off-diagonal blocks. Two off-diagonal blocks can clearly be discerned in the left panel of Figure 6. One of the blocks, marked by "A", indicates a correlation between the segments 1-80 and 420-550. The second off-diagonal block, marked by "B", indicates a correlation between the segments 80-350 and 550-787. Interestingly, this dependence structure is consistent with the results reported in Zhou and Spratt (1992). According to these authors, there are two "anomalous" regions in the DNA sequence alignment: at the beginning of the alignment, and in a segment around position 500. Now, block B captures the effect that most parts of the alignment are not "anomalous" - and hence related to each other. Block A indicates that the two "anomalous" regions, though different from the remainder of the alignment, are similar to each other. This is consistent with the findings in Zhou and

Spratt (1992), where both regions are reported as being "more diverged." Note that the MCP model, whose prediction is shown in the right panel of Figure 6, is oblivious to this pattern; in fact, its prediction is indistinguishable from one in which the two "anomalous" segments are unrelated (e.g. one being more and the other being less diverged).

6.4. Investigating the coupling of rates and transition-transversion ratios

Using the methods outlined in Section 5, we investigated if the model that couples the rate and transition-transversion ratio is favoured over the decoupled model. In this example, we set $\lambda_R = 1$, $\lambda_T = 1$, and $\lambda_J = 1$. The results were dependent on $\mathbf{C}$. If ν is not constrained, then the Bayes factor is 1.04 in favour of the coupled model, indicating that both models are approximately equally likely. However, if both models are focused on the interesting long scale behaviour, the Bayes factor is 1.90 in favour of the decoupled model, which the main model explored in this paper.

Our observation that the Bayes factor changes with a variation of the prior distribution of the parameters is well-known in Statistics. For example, for a vague prior distribution that poorly fits the data, the Bayes factor is known to favour the less complex model. This effect, which is related to Lindley's paradox, has e.g. been discussed in Jasra et al. (2005). We note that the situation is more complex in our situation, and can be explained biologically. As seen from the top right panel of Figure 2, there are two ranges of rate heterogeneity. The dominant peak on the left is related to the short-range effect of the genetic code. When constraining ν_R so as to exclude this effect, we allow the model to focus on long-range regional effects, related e.g. to different selective pressures, different codon biases etc. There is no biological reason why the rate heterogeneity and the transition-transversion ratio should be coupled here, and the Bayes factor, in fact, turns out to slightly favour the uncoupled chains. When not constraining ν_R, we mix the long-range behaviour with the short-range characteristics of the genetic code: the third codon position is well-known to be under less selective pressure due to the high prevalence of synonymous substitutions (meaning substitutions that do not change the amino acids). Interestingly, a recent study (Bofkin and Goldman, 2007) has shown that the transition-transversion ratio also follows this signature: First and second codon positions have virtually identical distributions of transition-transversion biases but the third codon position has a much higher average transition-transversion bias. Since both the overall rate variation, captured by the second chain, as well as the transition-transversion ratio, represented by the third chain, follow the same signature, the variations in the overall rate and the transition-transversion ratio are effectively coupled. By not constraining ν_R, we thus get a mixture of two effects: the short-range behaviour favouring the coupled, the long-range behaviour favouring the uncoupled model. This explains why, overall, no model is favoured and the Bayes factor approaches 1.

7. Segmentation of maize actin genes

We have applied our method to four maize actin gene sequences with the following GenBank/EMBL accession numbers: U60514, U60513, U60508, and U60507. We used the same alignment as in (Husmeier, 2005), and detected the same gene conversion event as found in this earlier study, in confirmation of the finding reported in Moniz de Sa and Drouin (1996). The new result we have obtained is shown in the left panel of Figure 7, which shows the posterior probability that two alignment positions are in the same rate state.

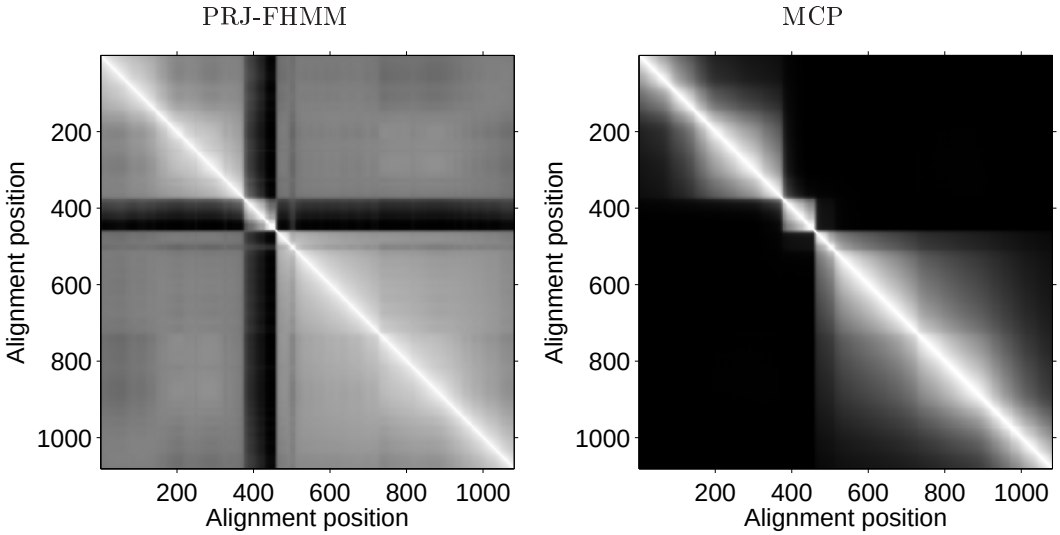

Figure 7. The posterior probability that two alignment positions in the maize actin gene sequence alignment are in the same rate state. The layout is identical to Figure 6.

Our method clearly detects an outlying region between sites 370 and 475. As it turns out (Lehrach, 2007), this region constitutes an inadvertently included intron, which owing to the lack of selective pressure is significantly more diverged than the coding regions by which it is flanked. Note that the corresponding plot obtained for the MCP model, shown in the right panel of Figure 7, is less clear in detecting this outlier, as it cannot be distinguished from an alignment of, say, three genes or three differently diverged coding regions. For further details, see Lehrach (2007).

8. Segmentation of a DNA sequence alignment of HIV-1

In 1996, a recombinant HIV-1 strain termed KAL153 caused an epidemic outbreak of AIDS infection among intravenous drug users around Kaliningrad, Russia. Liitsola et al. (1998) identified KAL153 as a recombinant of HIV-1 subtypes A and B. We analysed a whole genome sequence alignment of KAL153 with three consensus sequences of HIV-1 subtypes A, B, and F from the Los Alamos HIV Sequence Database. We used the same sequence alignment as in Suchard et al. (2003).

In the interest of space, we were advised by a referee to keep this section short. We therefore only summarise our main findings, and refer the reader to Lehrach (2007) for a comprehensive discussion of the results.

As we discussed before, the MCP model can be regarded as a special case of the PRJ-FHMM model. We should therefore be able to reproduce the results of Minin et al. (2005) by setting the PRJ-FHMM parameters to be equal to their respective counterparts in the MCP model. This was in fact confirmed in our study. Following Minin et al. (2005) we set $C_R^{\min} = 0.999$ and $C_T^{\min} = 0.999$, a lower bound equivalent to the setting of Minin et al. (2005) where $\lambda_R^B = 9$. Since HIV-1 contains ten major genes along the alignment (*gag*, *pro*, *pol*, *env*, *vif*, *vpr*, *vpu*, *tat*, *rev*, and *nef* – see Suchard et al., 2003), this corresponds to both

the MCP and the PRJ-FHMM expecting on average each gene to have its own rate and transition-transversion ratio. The results are shown in Figures 6.22 and 6.23 of Lehrach (2007), which confirm that the results obtained with the MCP and PRJ-FHMM models are indeed very similar.

The advantage of the proposed PRJ-FHMM model over the MCP model is that ν_R can be inferred from the sequence alignment. Hence rather than setting ν_R effectively fixed by a restrictive setting of $C_R^{\min}$ and $C_T^{\min}$, as described above for mimicking the setting of the MCP model, we relaxed the values of $C_R^{\min}$ and $C_T^{\min}$ and sampled ν_R from the posterior distribution with RJMCMC; see the discussion in Section 3. Surprisingly, we found a diffuse posterior distribution of ν_R that lacked the interpretable structure we had observed for Neisseria, as depicted in Figure 2. While on the face of it this looks like a failure, it in fact offers us a powerful diagnostic tool. Namely, the lack of any structure in the posterior distribution of ν_R indicates some fundamental problem with either the sequence alignment, or with some other aspect of the model, like an inappropriate nucleotide substitution model. If, for instance, the expected posterior distribution for ν_R is inferred when using an alternative alignment algorithm, it might indicate that this alternative alignment algorithm is superior. We feel that flagging these problems up is important for the user. This is an advantage over the MCP model, which does not infer the distribution of any parameter equivalent to ν_R and hence lacks the diagnostic potential to provide such indications.

9. Discussion

Boys et al. (2000) outlined some of the advantages that HMMs have compared to breakpoint models. When, for instance, recombination occurs in the middle of a DNA sequence alignment, the PRJ-FHMM can easily identify that the segments on either side have identical characteristics by assigning them to the same state, and thus with every extra occurrence of the state increases the confidence of estimating the state characteristics (see Lehrach (2007) for an example). In contrast, the MCP independently estimates the rate and transition-transversion ratio for each repeated occurrence of a state because it is modelled as a separate segment. The PRJ-FHMM can find repeated occurrences of states in a computationally efficient manner due to the existence of efficient algorithms for inference in HMMs.

Additionally, we have shown that the MCP of Minin et al. (2005) is effectively a special case of our model with a fixed value of ν_R, equivalent to the parameter λ_R^B, by Equation (24). In practice, there is uncertainty about how to set λ_R^B. The proposed Bayesian inference scheme addresses this uncertainty consistently by sampling ν_R from the posterior distribution. As seen from Figures 4 and 5, this leads to a considerable stabilisation of the predictions of the rate along the bacterial DNA alignment.

For the bacterial DNA sequence alignment analysed in Section 6, the PRJ-FHMM has provided independent verification for the claim of Zhou and Spratt (1992) that the anomalous region around $507 - 538$ is not the result of recombination. Instead, the PRJ-FHMM and the MCP have predicted that this anomalous region is part of a larger region of rate variation. In contrast to the MCP of Minin et al. (2005), we have also consistently identified rate variation occurring between alignment positions $1 - 75$, one of the more diverged regions found by Zhou and Spratt (1992). The MCP of Minin et al. (2005), on the other hand, did not consistently detect this region; the variability of its rate prediction results from the uncertainty in how to set λ_R^B. We have demonstrated that the proposed model infers the transition probability ν_R – equivalent to λ_R^B by Equation (24) – from the DNA

sequence alignment, picking up long scale behaviour, and codon position specific rate variation, and we have shown that by setting the hyperparameters $\mathbf{C}$ appropriately, we can focus on the behaviour we wish to investigate. Additionally, the decoupling of segments and states allowed us to predict that there were two to three different types of rate states present along the alignment.

For the HIV-1 DNA sequence alignment, we did not infer a clearly interpretable probability distribution of ν_R; this is in contrast to the distribution obtained for Neisseria, depicted in Figure 2b. A possible reason is inaccuracies in the DNA sequence alignment; owing to their large genetic diversification, HIV sequences are well-known to be intrinsically difficult to align. This points to an advantage of our proposed method over the MCP method of Minin et al. (2005), which does not include this inference step (λ_R^B, the parameter equivalent to ν_R, is set fixed) and hence lacks this diagnostic tool. However, when setting the value of ν_R to a fixed value corresponding to the value of λ_R^B used by Minin et al. (2005), our method effectively reproduces the authors' predictions.

10. Conclusions and future work

In this paper, we have investigated recombination and rate heterogeneity along three alignments: one of bacterial DNA sequences, another of HIV-1 DNA sequences, and a third of plant actin genes, where our analysis was performed using a fully Bayesian phylogenetic factorial hidden Markov model. The focus of our proposed model is simultaneously detecting recombination and characterising the rate states (or patterns of evolution) and rate segments of the alignments. This has many applications in functional genomics like identifying functional regions of proteins (Nimrod et al., 2005) and in comparative genomics (Chen and Blanchette, 2007) for detecting regulatory elements. In contrast, the focus of previous work such as Husmeier (2005) was on detecting recombination, and required choosing the set of possible rate states (or average branch lengths) in advance, leading to limited accuracy in characterising the rate, spurious predictions of rate variation (see Lehrach, 2007), and an inability to analyse the rate states.

We have shown that the MCP model of Minin et al. (2005) is a special case of our HMM formulation, and that generalising the model in a HMM formulation can produce results that the breakpoint model cannot obtain, such as posterior probabilities for different regions of the alignment sharing similar evolutionary characteristics. This is an additional benefit to consistently addressing the uncertainty inherent in the breakpoint model about how to set the number of rate segments. We also investigated if the rate and transition-transversion ratio should be estimated in a coupled fashion, and found that there was more support for these processes being separated when excluding short-range behaviour related to the signature of the genetic code. Both models are available in our software.

Our model can be enhanced to exploit the annotations that are available for the alignments, namely that given the positions of introns and exons, the codon effect can be modelled by defining codon position specific rate offsets. This was suggested by Felsenstein and Churchill (1996), but has not yet been integrated into our model and software.

There are many promising ways to exploit the model's ability to use considerably more complex state/segment specific evolutionary models, involving substantially more parameters than the HKY85 model employed in the present analysis. Consider using the model of Goldman and Yang (1994), which directly characterises the rate of nucleotide triplets with a detailed model involving 63 parameters; this will be almost unfeasible under a breakpoint

model as each segment would require independently estimating this large number of parameters (see also Kosiol et al. (2007) for a recent 71 parameter model). In contrast our proposed model would use all repeated occurrences of a state to estimate these parameters, due to its HMM formulation.

11. Acknowledgements

Wolfgang Lehrach and Dirk Husmeier are supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD). Wolfgang Lehrach was also supported by an ESPRC postgraduate student grant. We are grateful to two anonymous referees for very constructive feedback on an earlier version of our paper. Computing time was generously provided by Tony Travis on the Scottish Agricultural Research Institute (SARI) Linux Beowulf cluster.

References

Baldi, P. and P. Brunak (1998). *Bioinformatics - The Machine Learning Approach*. Cambridge, MA: MIT Press.

Bofkin, L. and N. Goldman (2007). Variation in Evolutionary Processes at Different Codon Positions. *Mol Biol Evol 24*(2), 513–521.

Boys, R. J. and D. A. Henderson (2001). A comparison of reversible jump MCMC algorithms for DNA sequence segmentation using hidden Markov models. *Computer Science and Statistics 33*, 35–49.

Boys, R. J. and D. A. Henderson (2004). A Bayesian approach to DNA sequence segmentation. *Biometrics 60*, 573–588.

Boys, R. J., D. A. Henderson, and D. J. Wilkinson (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics 49*, 269–285.

Casella, G. and E. I. George (1992). Explaining the Gibbs sampler. *The American Statistician 46*(3), 167–174.

Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*, 957–970.

Chen, H. and M. Blanchette (2007). Detecting non-coding selective pressure in coding regions. *BMC Evolutionary Biology 7*(Suppl 1), S9.

Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison (1998). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution 17*, 368–376.

Felsenstein, J. and G. A. Churchill (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution 13*(1), 93–104.

Gelman, A. and D. B. Rubin (1992, November). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

Ghahramani, Z. and M. I. Jordan (1997). Factorial hidden markov models. *Mach. Learn. 29*(2-3), 245–273.

Goldman, N. and Z. Yang (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol 11*(5), 725–736.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, 711–732.

Hasegawa, M., H. Kishino, and T. Yano (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution 22*, 160–174.

Husmeier, D. (2005). Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics 21*, ii166–ii172.

Husmeier, D., R. Dybowski, and S. Roberts (2005). *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Advanced Information and Knowledge Processing. New York: Springer.

Husmeier, D. and G. McGuire (2003). Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution 20*(3), 315–337.

Husmeier, D. and F. Wright (2001). Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology 8*(4), 401–427.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling.

Kosiol, C., I. Holmes, and N. Goldman (2007). An Empirical Codon Model for Protein Sequence Evolution. *Mol Biol Evol*.

Lehrach, W. P. (2007). *Predicting protein-protein interactions and characterising rate heterogeneity along DNA sequence alignments*. Ph. D. thesis, University of Edinburgh, School of Informatics.

Liitsola, K., I. Tashkinova, T. Laukkanen, G. Korovina, T. Smolskaja, O. Momot, N. Mashkilleyson, S. Chaplinskas, H. Brummer-Korvenkontio, J. Vanhatalo, P. Leinikki, and M. O. Salminen (1998). HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. *AIDS 12*, 1907–1919.

Maynard Smith, J. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution 34*, 126–129.

McGuire, G., F. Wright, and M. Prentice (2000). A Bayesian method for detecting past recombination events in DNA multiple alignments. *Journal of Computational Biology 7*(1/2), 159–170.

Minin, V. N., K. S. Dorman, F. Fang, and M. A. Suchard (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics 21*(13), 3034–3042.

Moniz de Sa, M. and G. Drouin (1996). Phylogeny and substitution rates of angiosperm actin genes. *Molecular Biology and Evolution 13*, 1198–1212.

Nimrod, G., F. Glaser, D. Steinberg, N. Ben-Tal, and T. Pupko (2005). In silico identification of functional regions in proteins. *Bioinformatics 21*, i328–337.

Pearl, J. (1988, September). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Rosenberg, M. S., S. Subramanian, and S. Kumar (2003). Patterns of transitional mutation biases within and among mammalian genomes. *Mol Biol Evol 20*(6), 988–993.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics 18*(3), 502–504.

Siepel, A. and D. Haussler (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology 11*(2-3), 413–428.

Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer (2003). Inferring spatial phylogenetic variation along nucleotide sequences: A multiple changepoint model. *Journal of the American Statistical Association 98*(462), 427–437.

Werhli, A., M. Grzegorczyk, M. Chiang, and D. Husmeier (2006). *Statistics in Genomics and Proteomics*, Chapter Improved Gibbs sampling for detecting mosaic structures in DNA sequence alignments, pp. 23–34. Centro Internacional de Matematica.

Zhou, J. and B. G. Spratt (1992). Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Molecular Microbiology 6*, 2135–2146.